日　本　国　特　許　庁
JAPAN　PATENT　OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application:　２００２年　６月２７日

出願番号
Application Number:　特願２００２－１８７６９８

[ST.10/C]:　[ＪＰ２００２－１８７６９８]

出願人
Applicant(s):　沖電気工業株式会社

２００３年　４月　１日

特許庁長官
Commissioner,
Japan Patent Office　太田信一郎

出証番号　出証特２００３－３０２２４４３

3175-1505-13

【書類名】　特許願
【整理番号】　KN002526
【提出日】　平成14年 6月27日
【あて先】　特許庁長官　及川　耕造　殿
【国際特許分類】　G06F 15/40
【発明者】
　【住所又は居所】　東京都港区虎ノ門１丁目７番１２号　沖電気工業株式会社内
　【氏名】　池野　篤司
【特許出願人】
　【識別番号】　000000295
　【氏名又は名称】　沖電気工業株式会社
　【代表者】　篠塚　勝正
【代理人】
　【識別番号】　100090620
　【弁理士】
　【氏名又は名称】　工藤　宣幸
【手数料の表示】
　【予納台帳番号】　013664
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9006358
【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　　情報分類装置

【特許請求の範囲】

【請求項１】　文書中の情報を分類する情報分類装置において、

入力文書をどのパターンを用いて分割、分類を行なうべきかを判定する文書種類判別部と、

表層的な分割パターンを用いて入力文書を分割する文書分割部と、

分割された部分文書に対して表層的なラベリングパターンを用いて適切な分類情報を付与するラベリング部と

を備えることを特徴とする情報分類装置。

【請求項２】　入力文書の表層的特徴から文書分割のパターンを自動生成する分割パターン生成部をさらに備えることを特徴とする請求項１に記載の情報分類装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、複数の情報が記載されている文書内の情報を分割して分類する情報分類装置に関するものである。

【０００２】

【従来の技術】

近年、インターネット等のネットワーク技術の普及により国内外の大量の電子文書へのアクセスが可能になり、大量の文書情報を分類する等の知的作業の自動化の必要性が高まってきている。

【０００３】

昨今発展を見せている電子文書の入手方法の一つに、メールマガジン（メールによる雑誌・新聞に類したもの）があげられる。これは、購読希望者に複数の情報をまとめて一つのメールに記載して送るというものである。

【０００４】

このようなメールは、複数の情報を記載した文書と見なすことができ、その情報を分類するためには文書内の各情報を適切に分割してやる必要がある。

【０００５】

特開２０００－２８５１４０号公報に記載の装置には、文書データの構造情報（ＨＴＭＬのタグや文字のフォント情報など）を基に文書データを分割する手段や、文書要素（例えば単語）や要素付随情報（例えば品詞）を基に文書データを分割する手段を設けることにより、情報の分類の一助としている例が示されている。

【０００６】

【発明が解決しようとする課題】

しかしながら、上記公報記載の装置では、メールマガジンのように明確な構造情報を持っていない文書には適用できないという問題がある。

【０００７】

また、仮に、あるメールマガジンを適切に分割する情報を指定したとしても、複数のメールマガジンを受け取っている場合、各々が異なる種類の分割情報（分割パターン）を必要とする可能性が高いが、メールマガジンの種類によって適切な分割パターンを選択して分割することができないという問題がある。

【０００８】

さらに、受け取るメールマガジンが増加すれば、分割パターンの種類も増加するが、それらを人手で指定するのは手間がかかるという問題がある。

【０００９】

以上の課題を解決するために、本発明は、複数の全く異なる情報が一つの文書内に含まれているもの、特にメールマガジン（例えばニュースのメールマガジン）のような、(i)構造情報は持たないが、人間が簡単に認識できるように記号などの表層情報を用いて明示的に内容の区切りが記述されている、(ii)文書種別によって区切りとなる表層情報や各部分の内容が分類するための情報が異なる、ことを特徴とするような文書を対象に、表層情報による分割パターンを使って分割し、分割後の各文書に分類情報を付与する情報分類装置を提供することを目的とする。

【００１０】

また、本発明は、種類の異なる分割パターンを持つ複数の文書が入力されるときに、自動的に最も適切な分割パターンを選択することができる情報分類装置を提供することを目的とする。

【００１１】

さらに、本発明は、システムが自動的に分割パターンを作成することができる情報分類装置を提供することを目的とする。

【００１２】

【課題を解決するための手段】

本発明は、文書中の情報を分類する情報分類装置において、入力文書をどのパターンを用いて分割、分類を行なうべきかを判定する文書種類判別部と、表層的な分割パターンを用いて入力文書を分割する文書分割部と、分割された部分文書に対して表層的なラベリングパターンを用いて適切な分類情報を付与するラベリング部とを備えることを特徴とする。

【００１３】

ここで、入力文書の表層的特徴から文書分割のパターンを自動生成する分割パターン生成部をさらに備えることが好ましい。

【００１４】

【発明の実施の形態】

（Ａ）第１の実施形態

以下、本発明による情報分類装置の第１の実施形態を図面を参照しながら詳述する。

【００１５】

（Ａ－１）第１の実施形態の構成

図１は、第１の実施形態の情報分類装置の機能的構成を示すブロック図である。例えば、第１の実施形態の情報分類装置は、通信機能を有するパソコン等の情報処理装置で実現されるが、機能的には、図１で表すことができる。

【００１６】

図１において、第１の実施形態の情報分類装置は、文書種類判別部１と、文書

分割部２と、ラベリング部３と、判別パターンデータ４と、分割パターンデータ５と、ラベリングパターンデータ６とを有する。

【００１７】

文書種類判別部１は、判別パターンデータ４を参照して、適用すべき分割パターンとラベリングパターンを決定するために、文書の種類を判別するものである。

【００１８】

文書分割部２は、文書種類判別部１の結果により、決定された分割パターンデータ５中のデータを適用して入力文書を分割するものである。

【００１９】

ラベリング部３は、文書種類判別部１の結果により、決定されたラベリングパターンデータ６中のデータを適用して、文書分割部２により分割された入力文書の各部分に対してラベリングを行なうものである。

【００２０】

判別パターンデータ４は、文書種類判別部１が文書の種別を判別するためのデータの集合である。最も単純な形式のデータとしては、特定の文字列（例えばメールマガジンであれば、マガジンのタイトルやＩＤ番号）があげられる。

【００２１】

図２に、判別パターンデータ４の一例を示す。各レコードは、文書種別と、その文書種類の適用する判別パターンとを含んでいる。

【００２２】

分割パターンデータ５は、文書分割部２が文書を分割するためのデータであり、例えば、図３に示すような文書種類と分割パターンとを対応付けたデータである。図３の分割パターンは、正規表現で記載されているので、パターン中の記号「＾」は「行頭」、「．」は「任意の一文字」、「＊」は「直前の文字が０回以上出現する」ことを意味している。よって「＾＝＝＝＝．＊」は「行頭から半角のイコール記号『＝』が４回出現した後にある文字が０回以上出現する」というパターンを示していることになる。

【００２３】

ラベリングパターンデータ６は、文書分割部２が分割した文書の各部分に対して、ラベリング部３が分類情報を付与する（ラベリングを行なう）ためのデータであり、図４に示すような、文書種類とラベリングパターンとラベル名とを対応付けたデータの集合である。図４に示すラベリングパターンも、正規表現で記載されている。

【００２４】

（Ａ－２）第１の実施形態の動作

以下、第１の実施形態の情報分類装置の動作を、各構成要素単位の動作で説明する。

【００２５】

まず、文書種類判別部１の動作を説明する。

【００２６】

文書種類判別部１は、判別パターンデータ４に収められた各パターンデータを用いて文書内をパターンマッチさせることで文書種類を判別するので、入力文書が図５のような文書であったときには、図２における第１番目のパターンデータの存在により、図５の文書は「ビジネスメールマガジン１」という種別であると判別される。

【００２７】

なお、複数のパターンデータがマッチし、かつ、その判別結果が矛盾する場合は、多数決により決定したり、矛盾が生じる旨をユーザに通知するなどの機能を設けてもよい。

【００２８】

次に、文書分割部２の動作を説明する。

【００２９】

文書分割部２は、上述したように、分割パターンデータ５に収められた各パターンデータを用いて、文書を分割する。図５の文書が「ビジネスメールマガジン１」という種別であり、図３のような分割パターンデータの例の場合、その種別に対して、図３の第１番目および第２番目の分割パターンが適用可能である。

【００３０】

すなわち、(i)先頭から「－」（半角のハイフン）が一定数以上連続している、(ii)先頭から「＝」（半角の等号）が一定数以上連続している、の部分が分割パターンとなるので、その位置で文書を部分文書に分割する。

【００３１】

分割後の各文書はデータ全般を記憶している記憶装置上に元データとは別に記憶されることになる。一般的な事項なので詳細は省略する。

【００３２】

また、分割に用いたパターンそのものの部分文書における取り扱いは、（１）分割後の部分文書には含まれない（パターンは削除される）、（２）分割位置の前後の部分文書のいずれかに含まれる、（３）分割位置の前後の両方の文書に含まれる（パターンは複製される）、のいずれかの方法で行うものとする。

【００３３】

（２）の方法により、図５の入力文書を分割した部分文書の例を図６に示している。

【００３４】

次に、ラベリング部３の動作を説明する。

【００３５】

ラベリング部３は、上述したように、ラベリングパターンデータ６に収められた各パターンデータを用いて、パターンがマッチした部分文書をラベリングする。図６のような部分文書群と、図４のようなラベリングパターンデータが存在した場合には、

部分文書１　　広告
部分文書２　　タイトル
部分文書３，４　　記事本文
部分文書５　　注釈

のようにラベリングされる。例えば、部分文書１には、「－－－ＰＲ－」というパターンが存在するので、図４の２番目の行が適用され、「広告」とラベリングされる。これらラベル情報は、各部分文書と組にして保持される。

【００３６】

（Ａ－３）第１の実施形態の効果

以上のように、第１の実施形態によれば、簡単なパターンによる分割パターンデータやラベリングパターンデータを用意するだけで、ＸＭＬやＨＴＭＬやＳＧＭＬで記述された明確な構造を持つ文書ではなくても、文書を分割して分類できるという効果がある。

【００３７】

また、文書種類判別部を設けたので、複数の分割パターンを管理しておき、様々な種類の文書を対象に文書を分割して分類できるという効果がある。

【００３８】

（Ｂ）第２の実施形態

次に、本発明による情報分類装置の第２の実施形態を図面を参照しながら詳述する。

【００３９】

（Ｂ－１）第２の実施形態の構成

図７は、第２の実施形態の情報分類装置の機能的構成を示すブロック図である。図示のように、本装置は、第１の実施形態の構成に、分割パターン生成部７を付加した構成となっている。

【００４０】

分割パターン生成部７は、入力文書を基に分割パターンを生成し、分割パターンデータ５に蓄積するものである。

【００４１】

これ以外の部分は、第１の実施形態と同様の機能を担っているので、その説明は省略する。

【００４２】

（Ｂ－２）第２の実施形態の動作

第１の実施形態と動作が異なるのは分割パターン生成部７だけなので、その動作のみを、図８のフローチャートを参照しながら説明する。

【００４３】

入力文書を受け取った後、ステップ８０１で入力文書を行ごとに分割する。次

に、先頭から複数文字（ここでは仮に３０文字とする）以上一致する行のグループを作る（ステップ８０２）。図５の入力文書を対象にグループを作成した結果の例を図９に示す。このうち、複数のメンバ（行）（ここでは２以上とする）を持つグループのみを選択してパターン記述を行う（ステップ８０３）。最も簡単なパターン記述法は文字列そのものであるが、必要に応じて正規表現などに書き改めるなどの手法を用いてもよいので、文書分割部２が理解できる形式を出力するものであれば特に手法は問わない。その後、分割パターンデータ５にデータを登録する（ステップ８０４）。

【００４４】

ステップ８０２の文字数のパラメータ、およびステップ８０３のメンバ（行）数のパラメータは自由に設定してもよい。また、ステップ８０２において「先頭から複数文字」としているが、「末尾から」であってもよいし、「先頭および末尾から」であっても「先頭・末尾関係なく」であってもよい。またそれらを自由に設定できる形式であってもよい。

【００４５】

（Ｂ－３）第２の実施形態の効果

以上のように、第２の実施形態によれば、システムが自動的に分割パターンを生成することができるという効果が、第１の実施形態の効果に追加される。

【００４６】

（Ｃ）他の実施形態

本発明においては、入力文書の分割および分割後の各部分に対するラベリングは同時に行なってもよい。

【００４７】

分割パターンデータをラベリングパターンデータの一部として用いる形式であってもよい。

【００４８】

【発明の効果】

以上のように、本発明によれば、複数の全く異なる情報が一つの文書内に含まれているような文書を対象に、表層情報による分割パターンを使って分割し、分

割後の各文書に分類情報を付与する情報分類装置を提供することができる。また、種類の異なる分割パターンを持つ複数の文書が入力されるときに、自動的に最も適切な分割パターンを選択することができる情報分類装置を提供することができる。

【図面の簡単な説明】

【図１】

第１の実施形態の情報分類装置の機能的構成を示すブロック図である。

【図２】

第１の実施形態の判別パターンデータ例を示す説明図である。

【図３】

第１の実施形態の分割パターンデータ例を示す説明図である。

【図４】

第１の実施形態のラベリングパターンデータ例を示す説明図である。

【図５】

第１の実施形態の動作説明に適用する入力文書例を示す説明図である。

【図６】

図５の入力文書に対する文書分割処理後のデータを示す説明図である。

【図７】

第２の実施形態の情報分類装置の機能的構成を示すブロック図である。

【図８】

第２の実施形態の分割パターン生成部の動作を示すフローチャートである。

【図９】

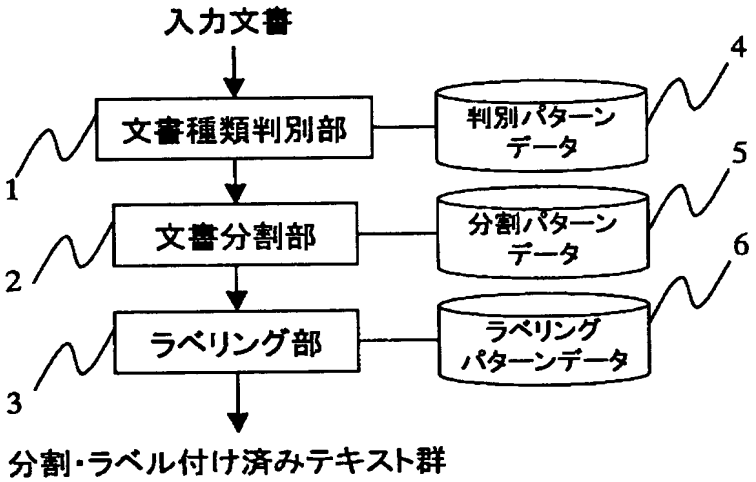
第２の実施形態の分割パターン生成時における入力文字のグループ化の説明図である。

【符号の説明】

１…文書種類判別部、２…文書分割部、３…ラベリング部、４…判別パターンデータ、５…分割パターンデータ、６…ラベリングパターンデータ、７…分割パターン生成部。

【書類名】　　図面

【図１】



【図２】

文書種類	判別パターン
ビジネスメールマガジン1	“DEFビジネスメールマガジン”
ビジネスメールマガジン1	ID=0000111
ビジネスメールマガジン2	“XYZニュース”
ビジネスメールマガジン2	MailMagazineID=999

【図３】

文書種類	分割パターン
ビジネスメールマガジン1	^====.*
ビジネスメールマガジン1	^----.*
ビジネスメールマガジン2	.....*
ビジネスメールマガジン2	^=-=-.*

出証特２００３－３０２２４４３

【図４】

文書種類	ラベリングパターン	ラベル
ビジネスメールマガジン1	^●.*	記事本文
ビジネスメールマガジン1	---PR-	広告
ビジネスメールマガジン1	ID=000.*	タイトル
ビジネスメールマガジン1	^Copyright.*	注釈
ビジネスメールマガジン2	^◆.*	記事本文
ビジネスメールマガジン2	^©.*	注釈

【図５】

---PR---
ABC Computer 新製品発売中！！

DEF ビジネスメールマガジン 朝刊
ID-0000111

●GHI株式会社、JKFコーポレーションとの提携を発表
http://www.aaa.bbb/ccc/ddd1234.html

GHI株式会社は××日、JKFコーポレーションと＊＊分野での提携に合意したと発表した。・・・
==========
●株式会社LMN、第1四半期の売上高は○○億円
http://www.aaa.bbb/ccc/ddd5678.html

LMNは△△の好調な売れ行きを反映して、第1四半期の売上高は○○億円と前期比＊＊%増を・・・
==========
Copyright 2002 OPQ社　無断転載を禁ず

【図６】

部分文書1
---PR----------
ABC Computer 新製品発売中！！

部分文書2

DEF ビジネスメールマガジン 朝刊
ID-0000111

部分文書3

●GHI株式会社、JKFコーポレーションとの提携を発表
http://www.aaa.bbb/ccc/ddd1234.html

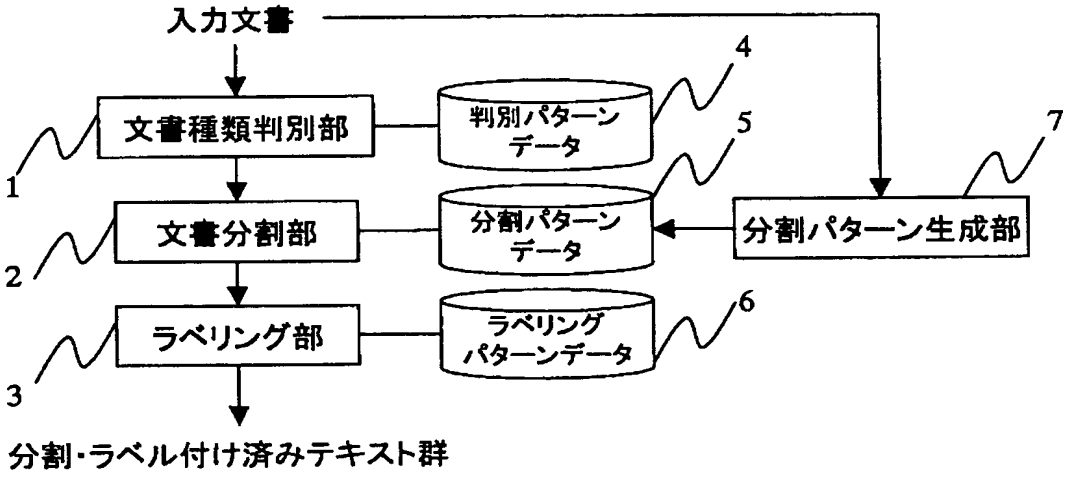
GHI株式会社は××日、JKFコーポレーションと＊＊分野での提携に合意したと発表した。・・・

部分文書4
==========
●株式会社LMN、第1四半期の売上高は○○億円
http://www.aaa.bbb/ccc/ddd5678.html

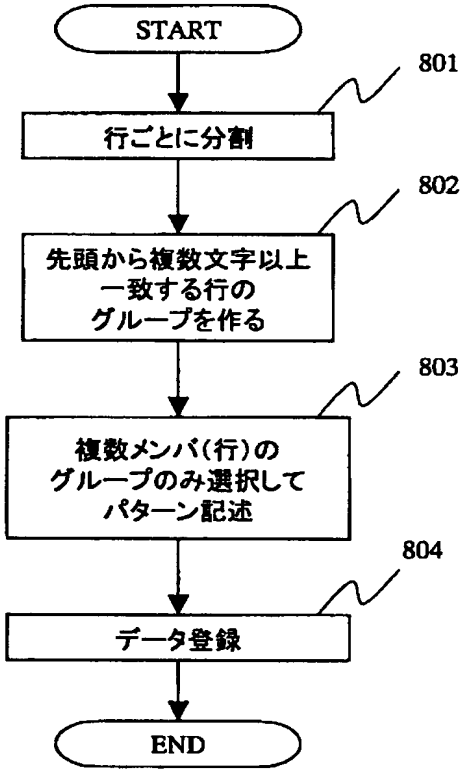
LMNは△△の好調な売れ行きを反映して、第1四半期の売上高は○○億円と前期比＊＊%増を・・・

部分文書5
==========
Copyright 2002 OPQ社　無断転載を禁ず

【図７】



【図８】



【図９】

先頭からの文字	行数
----(以下30個目まで)	2
==== (以下30個目まで)	2
---PR--- (以下30個目まで)	1
・・・	・・・
http://www.aaa.bbb/ccc/ddd1234	1
・・・	・・・

【書類名】　　要約書

【要約】

【課題】　複数の全く異なる情報が一つの文書内に含まれているような文書を分割し、分割後の各文書に分類情報を付与する情報分類装置を提供する。

【解決手段】　本発明は、文書中の情報を分類する情報分類装置に関する。そして、入力文書をどのパターンを用いて分割、分類を行なうべきかを判定する文書種類判別部と、表層的な分割パターンを用いて入力文書を分割する文書分割部と、分割された部分文書に対して表層的なラベリングパターンを用いて適切な分類情報を付与するラベリング部とを備えることを特徴とする。

ここで、入力文書の表層的特徴から文書分割のパターンを自動生成する分割パターン生成部をさらに備えることは好ましい。

【選択図】　　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００２９５〕

１．変更年月日　　１９９０年　８月２２日
［変更理由］　新規登録
住　所　　東京都港区虎ノ門１丁目７番１２号
氏　名　　沖電気工業株式会社